



INVESTOR IN PEOPLE

TI - DOCUMENT SUMMARIZING DEVICE, DOCUMENT SUMMARIZING METHOD AND RECORDING MEDIUM

PN - JP2001101228 A 20010413

AP - JP19990282041 19991001

OPD - 1999-10-01

PR - JP19990282041 19991001

PA - FUJI XEROX CO LTD

IN - KOYAMA TAKEHIRO

IC - G06F17/30

Best Available Copy

© WPI / DERWENT

TI - Document summarization apparatus has phrase synthesis section which synthesizes the phrase when inclusion of score in the new text is judged by a judging section

AB - JP2001101228 NOVELTY - A sentence analysis section (1a) analyses the sentence of input document and outputs analysis result. A specific portion of analysis result is converted into new text with respect to the data stored in the thesaurus (1d). A judging section (1e) judges whether a score is added to the new text. A phrase synthesis section synthesizes the phrase when inclusion of score in new text is judged.

- DETAILED DESCRIPTION - INDEPENDENT CLAIMS are also included for the following:

- (a) Document summarization procedure;

- (b) Recording medium

- USE - Document summarization apparatus.

- ADVANTAGE - The summary of the document which can be easily understandable, is generated.

- DESCRIPTION OF DRAWING(S) - The figure shows the block diagram of document summarization apparatus. (Drawing includes non-English language text).

- Sentence analysis section 1a

- Thesaurus 1d

- Judging section 1e

- (Dwg. 1/18)

PN - JP2001101228 A 20010413 DW200138 G06F17/30 017pp

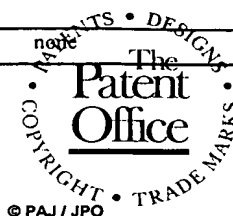
OPD - 1999-10-01

PR - JP19990282041 19991001

PA - (XERF) FUJI XEROX CO LTD

IC 06F17/30

AN - 2001-361456 [38]



INVESTOR IN PEOPLE

- TI** - DOCUMENT SUMMARIZING DEVICE, DOCUMENT SUMMARIZING METHOD AND RECORDING MEDIUM
- AB** - PROBLEM TO BE SOLVED: To generate a summary capable of easily grasping contents described in plural documents.
- SOLUTION: A parse tree of an inputted document is generated by a sentence analyzing part 1a, a partial tree is extracted from the parse tree by a partial tree creating part 1b, a new partial tree is generated by converting the partial tree by a partial tree conversion link part 1c, a pointer to indicate relation between the generated partial tree and its converting origin is imparted to the partial trees, scores are imparted to the partial trees, the partial trees are judged whether they are outputted as summaries or not in the order from the one with the highest score by a partial tree judging part 1e and the partial tree judged to be outputted is displayed by making it correspond to the partial tree of the converting origin of the partial tree judged to be outputted by a phrase synthesizing part 1f.
- PN** - JP2001101228 A 20010413
- AP** - JP19990282041 19991001
- PA** - FUJI XEROX CO LTD
- IN** - KOYAMA TAKEHIRO
- I** - G06F17/30

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-101228

(P2001-101228A)

(43) 公開日 平成13年4月13日 (2001.4.13)

(51) Int.Cl.⁷

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

15/40

テーマコード (参考)

3 2 0 A 5 B 0 7 5

3 7 0 A

審査請求 未請求 請求項の数 8 O L (全 17 頁)

(21) 出願番号

特願平11-282041

(22) 出願日

平成11年10月1日 (1999.10.1)

(71) 出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂二丁目17番22号

(72) 発明者 小山 剛弘

神奈川県足柄上郡中井町境430 グリーン

テクなかい 富士ゼロックス株式会社内

(74) 代理人 100092152

弁理士 服部 毅彦

Fターム (参考) 5B075 ND03 NK32 NK43 NK48 NS01

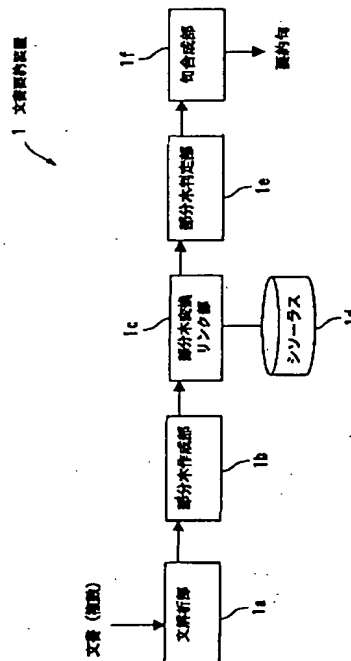
PQ02 PQ20 PQ69 PR04 UU05

(54) 【発明の名称】 文書要約装置、文書要約方法及び記録媒体

(57) 【要約】

【課題】 複数文書に記載された内容の把握を容易に行うことが可能な要約を生成する。

【解決手段】 文解析部1aによって、入力文書の解析木を生成し、部分木作成部1bによって、解析木から部分木を抽出し、部分木変換リンク部1cで部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポインタを付与し、部分木判定部1eによって、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否か判定し、句合成部1fによって、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示する。



【特許請求の範囲】

【請求項1】 入力された文書から要約を生成し、その出力を行う文書要約装置において、
 入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成する文解析手段と、
 前記文解析手段によって得られた前記解析木から部分木を生成する部分木作成手段と、
 前記部分木作成手段が生成した前記部分木を変換して新たな前記部分木を生成し、前記部分木と新たな前記部分木との関係を示すポインタを付与する部分木変換リンク手段と、
 前記部分木にスコアを付与し、前記スコアが高い順番に前記部分木を要約として出力するか否かを判定する部分木判定手段と、
 前記部分木判定手段が出力すると判定した前記部分木の要約を、出力すると判定された前記部分木の変換元の前記部分木の要約に対応付けて表示する句合成手段と、
 を有することを特徴とする文書要約装置。

【請求項2】 前記部分木変換リンク手段は、前記部分木が変換されたものでない場合、前記部分木の頻度を補正頻度とし、前記部分木が変換されたものである場合、前記部分木及び前記部分木の変換元に該当する部分木の頻度の合計数を前記補正頻度とし、
 前記部分木判定手段は、判定を行う前記部分木の前記補正頻度が、前記判定を行う前記部分木の変換元に該当する部分木の前記補正頻度以下の場合、前記部分木の出力を行わない判定を行うことを特徴とする請求項1記載の文書要約装置。

【請求項3】 前記部分木判定手段は、前記部分木が変換されたものでない場合には、前記補正頻度を前記スコアとし、前記部分木が変換されたものである場合には、前記補正頻度にスコア減率を乗じた値を前記スコアとすることを特徴とする請求項1記載の文書要約装置。

【請求項4】 用語の上位下位関係を記録した単語関係記録手段をさらに有し、前記部分木変換リンク手段は、前記単語関係記録手段の記録内容に従って前記部分木を前記部分木の上位語に変換して、新たな前記部分木を生成することを特徴とする請求項1記載の文書要約装置。

【請求項5】 前記部分木変換リンク手段は、前記部分木を複合語展開することによって、新たな前記部分木を生成することを特徴とする請求項1記載の文書要約装置。

【請求項6】 前記句合成手段は、まず、前記部分木変換リンク手段によって生成された新たな前記部分木及び自分から新たな前記部分木が生成されなかった前記部分木のみを表示し、次に、使用者からの表示指示によって、すべての前記部分木の表示を行うことを特徴とする請求項1記載の文書要約装置。

【請求項7】 入力された文書から要約を生成し、その

出力を行う文書要約方法において、

入力された文書からテキストを抽出し、
 抽出した前記テキストの係り受け解析を行うことによって解析木を生成し、

前記解析木から部分木を生成し、

前記部分木を変換して新たな前記部分木を生成し、

前記部分木と新たな前記部分木との関係を示すポインタを付与し、

前記部分木にスコアを付与し、

前記スコアが高い順番に前記部分木を要約として出力するか否かを判定し、

出力すると判定した前記部分木を、出力すると判定された前記部分木の変換元の前記部分木に対応付けて表示することを特徴とする文書要約方法。

【請求項8】 入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成し、

前記解析木から部分木を生成し、

前記部分木を変換して新たな前記部分木を生成し、

前記部分木と新たな前記部分木との関係を示すポインタを付与し、

前記部分木にスコアを付与し、

前記スコアが高い順番に前記部分木を要約として出力するか否かを判定し、

出力すると判定した前記部分木を、出力すると判定された前記部分木の変換元の前記部分木に対応付けて表示する機能をコンピュータに行わせるプログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は入力された文書の要約を生成する文書要約装置、文書要約方法及びその機能をコンピュータに行わせるプログラムを記録した記録媒体に関し、特に文書から抽出した部分木を合成することにより要約の生成を行う文書要約装置、文書要約方法及びその機能をコンピュータに行わせるプログラムを記録した記録媒体に関する。

【0002】

【従来の技術】これまで、複数の文書に記述されている内容の要約を生成する様々な文書要約方法が提案されてきた。このような文書要約方法としては、キーワード列挙 (Scatter/Gather 特開平5-225256)、複数文の合成 (特開平10-134066) 等が代表的なものである。キーワード列挙は、文書群を分類し、分類されたクラスターにそのクラスターで表れる頻出単語を列挙する方法であり、複数文の合成は、一連の文書群を、それらの冗長部分を削除して合成することにより要約を生成する方法である。

【0003】しかし、キーワード列挙の場合、抽出される結果は単なる単語の羅列となる為、単語間の関係が把

握できないという問題点がある。また、複数文の合成では適用文書の分野や形式が限定されるという問題がある。

【0004】このような問題を解決する文書要約方法として、本願出願人は、文書を係り受け解析して解析木を作成し、作成された解析木から部分木を切り出し、切り出した部分木をスコア加算しながら蓄積し、蓄積された部分木のスコアの値により、重要部分木を選択し、選択された重要部分木を合成することにより要約句を生成する複数文書要約装置の出願を行った（特願平10-358596）。

【0005】

【発明が解決しようとする課題】一般に、複数文書に使用される用語は、その形態が統一されていないことが多い為、部分木のスコア加算及び蓄積の際には、類義語展開や複合語展開による用語の変換を行う。しかし、特願平10-358596の方法では、変換によって用語が抽象化されすぎると、生成される要約の内容も抽象化されすぎることとなり、文書の内容の把握が困難になってしまうという問題点がある。例えば、用語が抽象化されすぎた場合、「A社が文書管理ソフトを発売」、「B社が文字認識ソフトを発売」、「C社が翻訳ソフトを発売」という文書から「会社がソフトを発売」という要約が作成されることもあり、この要約では内容が抽象化されすぎている為、文書の内容が十分把握できない。

【0006】また、抽象化が不適切な場合、生成される要約も不適切なものとなり、やはり文書の内容把握が困難になってしまうという問題点がある。さらに、特願平10-358596の方法では、類義語展開や複合語展開によって類似した部分木が多数作成され、その結果、類似した多数の要約が散乱して表示されるという問題点もある。

【0007】本発明はこのような点に鑑みてなされたものであり、複数文書に記載された内容の把握を容易に行うことが可能な要約を生成する文書要約装置を提供することを目的とする。

【0008】また、本発明の他の目的は、複数文書に記載された内容の把握を容易に行うことが可能な要約を生成することである。さらに、本発明の他の目的は、複数文書に記載された内容の把握を容易に行うことが可能な要約を生成する機能をコンピュータに行わせるプログラムを記録した記録媒体を提供することである。

【0009】

【課題を解決するための手段】本発明では上記課題を解決する為に、入力された文書から要約を生成し、その出力を行う文書要約装置において、入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成する文解析手段と、前記文解析手段によって得られた前記解析木から部分木を

生成する部分木作成手段と、前記部分木作成手段が生成した前記部分木を変換して新たな前記部分木を生成し、前記部分木と新たな前記部分木との関係を示すポイントを付与する部分木変換リンク手段と、前記部分木にスコアを付与し、前記スコアが高い順番に前記部分木を要約として出力するか否かを判定する部分木判定手段と、前記部分木判定手段が出力すると判定した前記部分木の要約を、出力すると判定された前記部分木の変換元の前記部分木の要約に対応付けて表示する句合成手段とを有することを特徴とする文書要約装置が提供される。

【0010】ここで、文解析手段は、入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成し、部分木作成手段は、文解析手段によって得られた前記解析木から部分木を生成し、部分木変換リンク手段は、部分木作成手段が生成した部分木を変換して新たな部分木を生成し、部分木と新たな部分木との関係を示すポイントを付与し、部分木判定手段は、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否かを判定し、句合成手段は、部分木判定手段が出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示する。

【0011】また、前記部分木変換リンク手段は、前記部分木が変換されていない場合、前記部分木の頻度を補正頻度とし、前記部分木が変換されている場合、前記部分木及び前記部分木の変換元に該当する部分木の頻度の合計数を前記補正頻度とし、前記部分木判定手段は、判定を行う前記部分木の前記補正頻度が、前記判定を行う前記部分木の変換元に該当する部分木の前記補正頻度よりも低い場合、前記部分木の出力を行わない判定を行うことを特徴とする文書要約装置が提供される。

【0012】これにより、抽象化が不適切な要約を排除することが可能となる。さらに、入力された文書から要約を生成し、その出力を行う文書要約方法において、入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成し、前記解析木から部分木を生成し、前記部分木を変換して新たな前記部分木を生成し、前記部分木と新たな前記部分木との関係を示すポイントを付与し、前記部分木にスコアを付与し、前記スコアが高い順番に前記部分木を要約として出力するか否かを判定し、出力すると判定した前記部分木を、出力すると判定された前記部分木の変換元の前記部分木に対応付けて表示することを特徴とする文書要約方法が提供される。

【0013】このようにすることにより、文書の内容把握を十分に行うことが可能な要約を生成することが可能となる。また、入力された文書からテキストを抽出し、抽出した前記テキストの係り受け解析を行うことによって解析木を生成し、前記解析木から部分木を生成し、前記部分木を変換して新たな前記部分木を生成し、前記部

分木と新たな前記部分木との関係を示すポイントを付与し、前記部分木にスコアを付与し、前記スコアが高い順番に前記部分木を要約として出力するか否か判定し、出力すると判定した前記部分木を、出力すると判定された前記部分木の変換元の前記部分木に対応付けて表示する機能をコンピュータに行わせるプログラムを記録した記録媒体が提供される。

【0014】このような記録媒体を用いることにより、文書の内容把握を十分に行うことが可能な要約生成をコンピュータに行わせることが可能となる。

【0015】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。まず、本発明における第1の実施形態について説明する。

【0016】図1に本形態における文書要約装置1の構成図を示す。文書要約装置1は、入力された文書からテキストを抽出し、抽出したテキストの係り受け解析を行うことによって解析木を生成する文解析手段である文解析部1a、解析木から部分木を生成する部分木作成手段である部分木作成部1b、部分木を変換して新たな部分木を生成し、部分木と新たな部分木との関係を示すポイントを付与する部分木変換リンク手段である部分木変換リンク部1c、用語の上位下位関係を記録した単語関係記録手段であるシソーラス1d、部分木を要約として出力するか否か判定する部分木判定手段である部分木判定部1e、及び要約として出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示する句合成手段である句合成部1fによって構成されている。

【0017】図2は、シソーラス1dに記録されているシソーラス辞書10の内容を示した図である。シソーラス辞書10には、複数の用語がそれらの上位下位関係とともに記録されており、いずれかの用語を選択することにより、その用語に対する同義語、上位語、兄弟語を抽出することが可能となっている。例えば、図2の場合、用語“きゅうり”を選択すると同義語は“胡瓜”、上位語は“野菜”及び“農作物”、兄弟語は“レタス”、“キャベツ”、“白菜”となる。

【0018】次に、図1及び図2を用いて、文書要約装置1の動作について説明する。まず、複数文書が文解析部1aに入力されると、文解析部1aは、入力された文書のテキストを抽出し、抽出したテキストの係り受け解析を行うことによって解析木を生成し、生成した解析木を部分木作成部1bに送る。

【0019】文解析部1aが生成した解析木を受け取った部分木作成部1bは、受け取った解析木から部分木を生成する。この際、生成された部分木を同じ部分木ごとに累積し、その累積数をその部分木に対する頻度とする。このように生成された部分木及びその頻度情報は、部分木変換リンク部1cに送られる。

【0020】部分木作成部1bによって生成された部分木及びその頻度情報を受け取った部分木変換リンク部1cは、受け取った部分木を変換し、新たな部分木の生成を行う。ここでの部分木の変換は、複合語変換及び類義語変換等により行う。複合語変換とは、複合語を後方一致の部分文字列に変換する方法であり、この方法により新たな部分木の生成を行う場合、部分木が有する複合語をその複合語の後方一致部分文字列に置き換える。例えば、複合語である用語“文書要約装置”を例に採ると、これに対する後方一致部分文字列は“要約装置”及び“装置”となり、この方法による新たな部分木の生成は、部分木が有する用語“文書要約装置”を“要約装置”及び“装置”に置き換えることによって行う。また、類義語変換とは、用語を同義語や上位語に変換する方法であり、この方法により新たな部分木の生成を行う場合、部分木が有する各用語を同義語や上位語に置き換える。例えば、用語“レタス”を例に採ると、シソーラス1dが有するシソーラス辞書10では、“レタス”の上位語は“野菜”及び“農作物”となり、この方法による新たな部分木の生成を行う場合、部分木変換リンク部1cはシソーラス1dを参照し、部分木が有する用語“レタス”を、その上位語である“野菜”及び“農作物”に置き換える。

【0021】新たな部分木の生成を行った部分木変換リンク部1cは、部分木に補正頻度を付与する。ここで補正頻度とは、部分木が部分木変換リンク部1cによって新たに生成されたものである場合には、部分木が有するすべての複合語変換元及び類義語変換元の部分木の頻度、及び新たに生成された部分木に等価な部分木の頻度を合計したものであり、部分木が新たに生成されたものでない場合には、その部分木の頻度を示すものとする。

【0022】また、部分木変換リンク部1cは、新たな部分木の生成を行う際、生成された新たな部分木にその変換元である部分木（リンク先）との関係を示すポイントを付与する。ポイントの付与は、新たに生成された部分木にリンク先の部分木の識別子を付与することにより行い、例えば、全ての部分木に番号を付与しておき、新たに生成された部分木にそのリンク先の部分木番号を付与することにより行う。ここで、部分木がリンク先を持たない場合、リンク先を持たないことを示すポイント“－1”等がその部分木に付与される。

【0023】このように生成された新たな部分木、部分木作成部1bで生成された部分木、それらに付与された頻度、補正頻度、及びポイントによって示されたリンク情報は、部分木判定部1eに送られる。

【0024】これらの情報を受け取った部分木判定部1eは、まず、各部分木に重要度を示すスコアを付与する。スコアリング方法はどのようなものでもよく、例えば、補正頻度をスコアとする方法、 $tf \times IDF$ 積 (G. Salton, “Automatic Text

Processing.” Addison-Wesley (1989) 等による単語重要度の和に、単語間の関係の種類ごとに与えた関係重要度を乗算する方法等を用いる。

【0025】スコアの付与を終了した部分木判定部1eは、部分木をそのスコアによってソートし、スコアの高い方から順番に、その部分木を要約として出力すべきか否かの判定を行う。ここで、判定を行った部分木がリンク先を有する場合、続けてリンク先の部分木の判定も行い、一旦判定が行われた部分木については、再び判定が行われることはない。このように判定された判定結果は判定情報として各部分木に付与される。判定情報の付与は、例えば、判定前の部分木の判定情報を0で初期化しておき、要約として出力する部分木には1の判定を、出力しないと判定した部分木には-1の判定を、判定1が付与された部分木のリンク先の部分木について、要約として出力する場合には、そのリンク先の部分木に2の判定を、出力しない場合には-1の判定をそれぞれ該当する部分木に付与することにより行う。

【0026】次に、フローチャートを用いて、部分木判定部1eが行う判定手順例の説明を行う。図3及び図4は、部分木判定部1eが行う判定の判定手順を例示したフローチャートである。

【0027】〔S1〕 部分木がリンク先を有しているか否かを判断する。リンク先を有している場合には、ステップS3に進む。リンク先を有していない場合には、ステップS2に進む。

【0028】〔S2〕 要約を出力するとして、その部分木に判定1を設定する。

〔S3〕 判定した部分木以上の補正頻度を有するリンク先が存在するか否かを判断する。存在する場合には、ステップS5に進む。存在しない場合には、ステップS4に進む。

【0029】〔S4〕 判定1のリンク先が存在するか否かを判断する。存在する場合には、ステップS5に進む。存在しない場合には、ステップS6に進む。

〔S5〕 要約を出力しないとして、その部分木に判定-1を設定する。

【0030】〔S6〕 要約を出力するとして、その部分木に判定1を設定する。

〔S7〕 リンク先の部分木の判定を行う。

図4は、図3のステップS7の詳細を示したフローチャートである。

【0031】〔S11〕 部分木がさらにリンク先を有しているか否かを判断する。リンク先を有している場合には、ステップS13に進む。リンク先を有していない場合には、ステップS12に進む。

【0032】〔S12〕 要約を出力するとして、その部分木に判定2を設定する。

〔S13〕 判定した部分木以上の補正頻度を有するリ

ンク先が存在するか否かを判断する。存在する場合には、ステップS15に進む。存在しない場合には、ステップS14に進む。

【0033】〔S14〕 判定1のリンク先が存在するか否かを判断する。存在する場合には、ステップS15に進む。存在しない場合には、ステップS16に進む。

〔S15〕 要約を出力しないとして、その部分木に判定-1を設定する。

【0034】〔S16〕 要約を出力するとして、その部分木に判定2を設定する。

〔S17〕 ステップS11からステップS16までの手順を再帰的に適用し、さらなるリンク先の判定を行う。

【0035】このように判定が行われた各部分木、それらの頻度、補正頻度、ポイントによって示されたリンク情報、スコア及び判定結果の情報は、句合成部1fに送られる。

【0036】これらの情報を受け取った句合成部1fは、受け取った情報を元に要約を合成し、その表示を行う。要約の合成は、スコアが高い部分木から順番に、図5に示すフローチャートに従って行われる。

【0037】〔S21〕 部分木に付与された判定結果を判断する。判定が1である時はステップS22に進む。判定が-1或いは2である時はステップS23に進む。

〔S22〕 ステップS21で判断を行った部分木によって要約を合成し、その要約を部分木の補正頻度とともに表示する。

【0038】〔S23〕 ステップS21で判断を行った部分木をスキップする。

〔S24〕 ステップS21で判断を行った部分木がリンク先を有するか否かを判断する。リンク先を有する場合、ステップS25に進む。リンク先を有しない場合、処理を終了する。

【0039】〔S25〕 ステップS21で判断を行った部分木の頻度が1以上か否かを判断する。1以上である場合、ステップS26に進む。0である場合、ステップS27に進む。

【0040】〔S26〕 ステップS21で判断を行った部分木によって要約を合成し、その要約を部分木の頻度とともに、ステップS22で表示された要約に関連付けて表示する。

【0041】〔S27〕 リンク先の部分木の判定が2で頻度が1以上であるか否かを判断する。リンク先の部分木の判定が2で頻度が1以上である場合、ステップS28に進む。そうでない場合、処理を終了する。

【0042】〔S28〕 リンク先の部分木によって要約を合成し、その要約を部分木の頻度とともに、ステップS22で表示された要約に関連付けて表示する。なお、図5のフローチャートではステップS26及びステ

ップS28において、ステップS21で判断を行った部分木によって合成された要約(ステップS26)、及びリンク先の部分木によって合成された要約(ステップS28)を、ステップS22で表示された要約に従属させて表示することとしたが、ステップS26及びステップS28での要約の表示を、使用者からの表示指示があった場合にのみ行うこととしてもよい。

【0043】次に、実際の文書例を用いて文書要約装置1の動作の説明を行う。図6は、複数文書2“レタスの価格が高い…”、“キャベツの価格が高い…”、“白菜の価格が高い…”を入力文書とした場合における文書要約装置1の要約生成過程を示した図である。なお、各文書は複数文から構成されているが、ここでは説明を簡単にする為に1文に注目して説明する。

【0044】文解析部1aは、複数文書2からテキスト情報(レタスの価格が高い。キャベツの価格が高い。白菜の価格が高い。)を抽出し、それらの係り受け解析を行い、解析木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)を生成する。ここで、解析木(レタス-の-価格-が-高い)は、“レタス”が“価格”に関係“の”で係り、“価格”が“高い”に関係“が”で係っていることを表わす。生成された解析木は、解析木情報として部分木作成部1bに送られる。

【0045】解析木情報を受け取った部分木作成部1bは、受け取った解析木から部分木を生成する。ここで、既に、同じ部分木が作成されている時は、頻度のみを加算する。部分木は、解析木のサブツリーとして複数作成されるが、ここでは便宜上解析木そのものを部分木とする。このように生成された部分木の集合である部分木情報2a(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)、及びそれらに対応する頻度2bは、部分木変換リンク部1cに送られる。ここで、部分木情報2aが有する各部分木の頻度2bは全て1となる。

【0046】部分木情報2aを受け取った部分木変換リンク部1cは、部分木情報2aが有する部分木を変換し、新しい部分木の生成を行う。ここでは、類義語変換によって新しい部分木の生成を行うものとする。部分木変換リンク部1cは、部分木情報2a(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)を得ると、シソーラス辞書10を参照して部分木の変換を行い、新たな部分木の生成を行う。ここで、シソーラス辞書10には、用語“レタス”、“キャベツ”、“白菜”の上位語として“野菜”が記録されている為、“レタス”、“キャベツ”、“白菜”を“野菜”に変換した新たな部分木(野菜-の-価格-が-高い)が生成される。さらに、“野菜”の上位語として“農作物”が存在する為、新たな部分木(野菜-の-価格-が-高い)の“野菜”を“農作物”に変換

した新たな部分木(農作物-の-価格-が-高い)が生成される。図6では、このように生成された新たな部分木(野菜-の-価格-が-高い、農作物-の-価格-が-高い)、及び部分木情報2aに含まれる部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)の集合を変換部分木情報2cとして示している。ここで、新たに生成された部分木(野菜-の-価格-が-高い、農作物-の-価格-が-高い)は、部分木情報2aには含まれていないものである為、これらの頻度2bは0となる。

【0047】次に、部分木変換リンク部1cは、変換部分木情報2cに補正頻度2dを付与する。ここで、部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)は、部分木変換リンク部1cによって新たに生成されたものでない為、それらの頻度2bの値1をそのまま補正頻度2dの値とする。また、部分木(野菜-の-価格-が-高い)は、部分木変換リンク部1cによって新たに生成されたものである為、その変換元の部分木に該当する部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)の頻度2bの合計数3をそれぞれの補正頻度2dとする。また、部分木(農作物-の-価格-が-高い)も新たに生成されたものである為、その変換元の部分木に該当する部分木(野菜-の-価格-が-高い)及び(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)の頻度の合計数3を、部分木(農作物-の-価格-が-高い)の補正頻度2dとする。

【0048】また、部分木変換リンク部1cは、変換部分木情報2cのリンク先との関係を示すリンク情報であるボインタの付与を行う。ここで、部分木(レタス-の-価格-が-高い)の部分木番号を1、部分木(キャベツ-の-価格-が-高い)の部分木番号を2、部分木(白菜-の-価格-が-高い)の部分木番号を3、部分木(野菜-の-価格-が-高い)の部分木番号を4、部分木(農作物-の-価格-が-高い)の部分木番号を5とし、ボインタの付与を行う。

【0049】部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)はリンク先を持たない為、これらのリンク2eに示されるボインタは、リンク先がないことを表す-1となる。また、部分木(野菜-の-価格-が-高い)は、リンク先として部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)を持つ為、このリンク2eに示されるボインタは、部分木(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)の部分木番号である1、2、3となる。また、部分木(農作物-の-価格-が-高い)は、リンク先として部分木(野菜-の-価格-が-高い)を持つ為、このリンク2eに示

されるポイントは、部分木（野菜 - の - 価格 - が - 高い）の部分木番号である4となる。

【0050】このように部分木変換リンク部1cによって生成された結果を図7に示す。図7に示した生成結果は、部分木判定部1eに送られる。これらの情報を受け取った部分木判定部1eは、まず、変換部分木情報2cのスコア2fの付与を行う。ここでは簡単にする為、補正頻度2dの値をスコア2fの値とする。

【0051】次に、部分木判定部1eは、変換部分木情報2cをスコア2fによってソートし、スコアの高い変換部分木情報2cから順番に、その変換部分木情報2cを要約として出力するか否かの判定を行う。

【0052】この例の場合、最もスコアの高い部分木は、部分木番号4の（野菜 - の - 価格 - が - 高い）であり、この部分木はリンク先として部分木番号1、2、3の部分木を有している（ステップS1）。また、リンク先である部分木番号1、2、3の部分木の補正頻度は部分木番号4の部分木の補正頻度よりも小さく（ステップS3）、さらに、リンク先である部分木番号1、2、3の部分木の中には、判定1のものが存在しない（ステップS4）。以上より、部分木番号4の部分木（野菜 - の - 価格 - が - 高い）を要約として出力するものとして、部分木番号4の部分木の判定2gを1とする（ステップS6）。

【0053】次に、部分木番号4の部分木のリンク先である部分木番号1、2、3の部分木の判定を行う（ステップS7）。リンク先である部分木番号1の部分木（レタス - の - 価格 - が - 高い）、部分木番号2の部分木（キャベツ - の - 価格 - が - 高い）、部分木番号3の部分木（白菜 - の - 価格 - が - 高い）は、リンク先を有していない為（ステップS11）、部分木番号1、2、3の部分木を要約として出力するものとして、部分木番号1、2、3の部分木の判定2gを2とする（ステップS12）。

【0054】部分木番号4の部分木の判定が終了すると、次に、同じ値のスコア2fを持つ部分木番号5の部分木の判定を行う。部分木番号5の部分木（農作物 - の - 価格 - が - 高い）は、部分木番号4の部分木をリンク先として持ち（ステップS1）、リンク先である部分木番号4の部分木は、部分木番号5の部分木以上の補正頻度2d（3）を持つ為（ステップS3）、部分木番号5の部分木を要約として出力しないものとして、部分木番号5の部分木の判定2gを-1とする（ステップS5）。

【0055】次にスコアの高い部分木番号1、2、3の部分木については、判定2gが既に2に設定されている為、チェックを行わない。以上により、図8に示すような判定情報が生成される。部分木判定部1eは、このように生成された判定情報を句合成部1fに送る。

【0056】判定情報を受け取った句合成部1fは、図

5のフローチャートに従って、要約の合成を行う。まず、部分木番号4の部分木（野菜 - の - 価格 - が - 高い）の判定2gは1である為（ステップS21）、要約「野菜の価格が高い」を合成し、その補正頻度2dである[3]とともに表示する（ステップS22）。部分木番号4の部分木はリンク先を有し（ステップS24）、部分木番号4の部分木の頻度2bは0であり（ステップS25）、リンク先である部分木番号1、2、3の部分木の判定2gが2、頻度2bが1以上である為（ステップS27）、要約「レタスの価格が高い、キャベツの価格が高い、白菜の価格が高い」を合成し、それらの頻度2bである[1]とともに、部分木番号4の部分木から生成された要約「野菜の価格が高い」に関連付けて表示する（ステップS28）。

【0057】部分木番号5の部分木は判定2gが-1なのでスキップし、部分木番号1、2、3の部分木は判定2gが2なのでスキップする。以上より、図9の（a）に示す要約が生成される。なお、ここでは、段落をつけて関連する要約を示しているが、ツリー構造など他の方法で関連する要約を表示してもよい。また、図9の（b）に示すように、まず、判定2gが1である部分木から生成した要約のみを表示し、ユーザの指示によって、そのリンク先の要約を表示することとしてもよい。

【0058】このように、本形態では、文解析部1aによって、入力文書の解析木を生成し、部分木作成部1bによって、解析木から部分木を抽出し、部分木変換リンク部1cで部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木判定部1eによって、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否か判定し、句合成部1fによって、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示することとした為、抽象度の高い「農作物の価格が高い」が排除され、最適な抽象度の「野菜の価格が高い」が要約として選択され、選択された「野菜の価格が高い」が元の表現「レタスの価格が高い、キャベツの価格が高い、白菜の価格が高い」と関連させて表示されることとなり、ユーザはこの表示から野菜全般の価格が高いのではなく、「レタス、キャベツ、白菜」といった「葉もの野菜」の価格が高いことを判断することができる。

【0059】次に、本形態における第2の実施の形態について説明する。本形態は、第1の実施の形態の文書要約装置1において、複合語展開によって新たな部分木を生成する場合を示した例である。図10は、複数文書3「野菜の価格が高い…」、「野菜の価格の全国調査…」、「農作物の価格を調査する…」、「野菜の価格が高い…」を入力文書とした場合における文書要約装置1の要約生成過程を示した図である。なお、各文書は複数文から構成されているが、ここでは説明を簡単にする為

に1文に注目して説明する。また、文解析部1a及び部分木作成部1bにおける動作は、第1の実施の形態と同様である為、説明を省略する。

【0060】部分木変換リンク部1cは、部分木情報3a（野菜-の-価格-が-高い、野菜-の-価格-の-全国調査、農作物-の-価格-を-調査）を得ると、シソーラス1dが有するシソーラス辞書10を参照して部分木の変換を行う。なお、ここでは、部分木（野菜-の-価格-が-高い）の部分木番号を1、部分木（野菜-の-価格-の-全国調査）の部分木番号を2、（農作物-の-価格-を-調査）の部分木番号を3として説明を行う。また、ここでの部分木番号1の部分木の頻度3bは2、部分木番号2の部分木の頻度3bは1、部分木番号3の部分木の頻度3bは1となる。

【0061】部分木変換リンク部1cは、部分木番号1の部分木（野菜-の-価格-が-高い）の用語“野菜”を“農作物”に変換した新たな部分木（農作物-の-価格-が-高い）を追加し、その部分木番号を4、頻度3bを0、補正頻度3dを2として設定する。次に、部分木番号2の部分木（野菜-の-価格-の-全国調査）において、用語“野菜”を“農作物”に変換し、用語“全国調査”を複合語展開によって“調査”に変換することにより生成される新たな部分木（農作物-の-価格-の-全国調査、野菜-の-価格-の-調査）を、それぞれ部分木番号5及び6として追加する。ここで、同様に生成される部分木（農作物-の-価格-の-調査）は、部分木番号3の部分木と等価であるので、この変換による新たな部分木は生成せず、部分木番号3の補正頻度3dのみを追加する。部分木番号3の部分木（農作物-の-価格-を-調査）については複合語、類義語を有していない為変換は行われない。

【0062】部分木番号1、2の部分木はリンク先を有していない為、これらの補正頻度3dは、これらの頻度3bの値をそのまま用い、2及び1として設定する。また、部分木番号3の部分木は、部分木番号2の部分木から変換された部分木と等価である為、部分木番号2の部分木及び部分木番号3の部分木の頻度3bを合計した値2を部分木番号3の部分木の補正頻度3dとする。また、部分木番号4、5、6の部分木は、部分木番号1、2の部分木からそれぞれ変換されたものである為、部分木番号1、2の頻度3bを部分木番号4、5、6の部分木の補正頻度3dとする。これにより、部分木番号4の部分木の補正頻度3dを2、部分木番号5、6の部分木の補正頻度3dをそれぞれ1として設定することとなる。

【0063】また、部分木番号1、2の部分木はリンク先を有していない為、これらのリンク3eの値は-1となり、部分木番号3、5、6の部分木は部分木番号2の部分木をリンク先として有している為、これらのリンク3eの値は2となり、同様に、部分木番号4の部分木のリ

ンク3eの値は、そのリンク先の部分木番号1となる。

【0064】このように部分木変換リンク部1cによって生成された結果を図11に示す。図11に示した生成結果は、部分木判定部1eに送られる。部分木判定部1eは、図11に示した生成結果を得ると各部分木にスコアを付与し、スコアでソートを行い、スコアが高い部分木から順番に要約を出力すべきかどうかの判定を行う。補正頻度をスコアとすると、最もスコアの高い部分木は部分木番号1の部分木（野菜-の-価格-が-高い）となり、部分木番号1の部分木はリンク先を有していない為（ステップS1）、要約を出力するとして判定3gを1とする（ステップS2）。

【0065】次に、部分木番号1の部分木と同じスコアを持つ部分木番号3の部分木（農作物-の-価格-を-調査）の判定を行う。部分木番号3の部分木はリンク先として部分木番号2の部分木を持ち（ステップS1）、リンク先である部分木番号2の部分木の補正頻度3b（1）が、部分木番号3の部分木の補正頻度3b（2）以上ではなく（ステップS3）、判定3gも1ではない為（ステップS4）、部分木番号3の部分木の判定3gを1として設定する（ステップS6）。

【0066】次に、部分木番号3の部分木のリンク先に該当する部分木の判定を行う（ステップS7）。リンク先である部分木番号2の部分木（野菜-の-価格-の-全国調査）は、リンク先を有していない為（ステップS11）、部分木番号2の部分木の判定3gを2として設定する（ステップS12）。

【0067】次に、部分木番号3の部分木と同じスコアを持つ部分木番号4の部分木（農作物-の-価格-が-高い）の判定を行う。部分木番号4の部分木はリンク先として部分木番号1の部分木を有し（ステップS1）、リンク先である部分木番号1の部分木の補正頻度3d（2）が部分木番号4の部分木の補正頻度3d（2）以上である為（ステップS3）、部分木番号4の部分木の判定3gを-1として設定する（ステップS5）。

【0068】次にスコアが高い部分木番号2の部分木（野菜-の-価格-の-全国調査）は、判定3gが既に2に設定されている為チェックを行わない。次に、部分木番号5、6部分木（農作物-の-価格-の-全国調査、野菜-の-価格-の-調査）は、リンク先を有し（ステップS1）、リンク先の部分木の補正頻度3d（1）が自分自身の補正頻度3d（1）以上なので（ステップS3）、判定3gをそれぞれ-1として設定する（ステップS5）。

【0069】以上により、図12に示すような判定情報が生成される。部分木判定部1eは、このように生成された判定情報を句合成部1fに送る。判定情報を受け取った句合成部1fは、図5のフローチャートに従って、要約の合成を行う。

【0070】まず、部分木番号1の部分木の判定3gは

1である為(ステップS21)、部分木番号1の部分木より合成した要約「野菜の価格が高い」を生成し、生成した要約を、部分木番号1の部分木における補正頻度3dの値[2]とともに表示する(ステップS22)。部分木番号1の部分木はリンク先を有していない為(ステップS24)、次の部分木の表示に移る。

【0071】次に、部分木番号3の部分木の判定3gは1である為(ステップS21)、部分木番号3の部分木より合成した要約「農作物の価格を調査」を生成し、生成した要約を、部分木番号3の部分木における補正頻度3dの値[2]とともに表示する(ステップS22)。部分木番号3の部分木は、リンク先として部分木番号2の部分木を有し(ステップS24)、自分自身の頻度3bが1である為(ステップS25)、部分木番号3の部分木から要約「農作物の価格を調査」を生成し、生成した要約を、その頻度3bの値[1]とともに、ステップS22で表示された要約「農作物の価格を調査」に関連付けて表示する(ステップS26)。また、リンク先の部分木である部分木番号2の部分木の判定3gが2であり、その頻度が1である為(ステップS27)、部分木番号2の部分木から要約「野菜の価格の全国調査」を生成し、その頻度3bの値[1]とともにステップS22で表示された要約「農作物の価格を調査」に関連付けて表示する(ステップS28)。

【0072】部分木番号4の部分木は判定3gが-1なのでスキップし、部分木番号2の部分木は判定3gが2なのでスキップし、部分木番号5、6の部分木は判定3gが-1なのでスキップする。

【0073】以上より、図13の(a)に示す要約が生成される。なお、ここでは、段落をつけて関連する要約を示しているが、ツリー構造など他の方法で関連する要約を表示してもよい。また、図13の(b)に示すように、まず、判定3gが1である部分木から生成した要約のみを表示し、ユーザの指示によって、そのリンク先の要約を表示することとしてもよい。

【0074】このように本形態では、文解析部1aによって、入力文書の解析木を生成し、部分木作成部1bによって、解析木から部分木を抽出し、部分木変換リンク部1cで部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木判定部1eによって、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否か判定し、句合成部1fによって、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示することとした為、頻度3bとリンク3eから不要なデータ(部分木番号4、5、6の部分木)を排除し、関連する要約をまとめて表示することができる。

【0075】次に、本発明における第3の実施の形態について説明する。本形態は、第1の実施の形態の文書要

約装置1において、部分木のスコアを付与する際に部分木の重みを考慮してスコアリングを行った場合を示した例である。図14は、複数文書4“レタスの価格が高い…”×10、“キャベツの価格が高い…”、“白菜の価格が高い…”を入力文書とした場合における文書要約装置1の要約生成を示した図である。なお、各文書は複数文から構成されているが、ここでは説明を簡単にする為に1文に注目して説明する。また、文解析部1a及び部分木作成部1bにおける動作は、第1の実施の形態と同様である為、説明を省略する。

【0076】部分木変換リンク部1cは、部分木情報4a(レタス-の-価格-が-高い、キャベツ-の-価格-が-高い、白菜-の-価格-が-高い)を得ると、シソーラス1dが有するシソーラス辞書10を参照して部分木の変換を行う。なお、ここでは、部分木(レタス-の-価格-が-高い)の部分木番号を1、部分木(キャベツ-の-価格-が-高い)の部分木番号を2、(白菜-の-価格-が-高い)の部分木番号を3として説明を行う。また、ここでの部分木番号1の部分木の頻度4bは10、部分木番号2の部分木の頻度4bは1、部分木番号3の部分木の頻度4bは1となる。

【0077】部分木変換リンク部1cは、シソーラス1dが有するシソーラス辞書10を参照し、部分木番号1の部分木(レタス-の-価格-が-高い)の用語“レタス”を上位語“野菜、農作物”に変換した部分木番号4、5の部分木(野菜-の-価格-が-高い、農作物-の-価格-が-高い)を追加し、それらの頻度4bを0、補正頻度4dを12とする。そして、部分木番号4の部分木(野菜-の-価格-が-高い)のリンク4eにリンク先である部分木の部分木番号1、2、3を付与し、部分木番号5の部分木(農作物-の-価格-が-高い)のリンク4eにリンク先である部分木の部分木番号4を付与する。

【0078】同様に、部分木番号2の部分木(キャベツ-の-価格-が-高い)、部分木番号3の部分木(白菜-の-価格-が-高い)から“キャベツ、白菜”を上位語“野菜、農作物”に変換した部分木(野菜-の-価格-が-高い、農作物-の-価格-が-高い)が生成されるが、生成されたこれらの部分木は既に生成された部分木番号4、5の部分木と等価である為、新たな部分木は生成せず、部分木番号2の部分木における頻度4bの値1を、部分木番号4、5の部分木における補正頻度4dの値に加算する。また、部分木番号4の部分木のリンク4eに、リンク先となる部分木の部分木番号2、3を追加する。

【0079】このように部分木変換リンク部1cによって生成された結果を図15に示す。図15に示した生成結果は、部分木判定部1eに送られる。これらの情報を受け取った部分木判定部1eは、まず、変換部分木情報4cのスコア4fの付与を行う。スコアの与え方は色々

考えられるが、ここでは基本的に補正頻度4 dをスコア4 fとし、部分木が複合語展開や類義語展開により抽象化されたものである場合には、その部分木の補正頻度4 dにスコア減減率(0~1)を乗じたものをスコア4 fとする。

【0080】図16に、スコア減減率を0.7に設定した場合の結果を示す。ここで、部分木番号1、2、3の部分木は抽象化されたものではないので、それらの補正頻度4 dをスコア4 fとし、部分木番号4の部分木(野菜-の-価格-が-高い)は、部分木番号1、2、3の部分木が有する用語“レタス、キャベツ、白菜”を上位語“野菜”に変換することにより生成したものであるため、そのスコア4 fは $12 \times 0.7 = 8.4$ となり、部分木番号5の部分木(農作物-の-価格-が-高い)は、部分木番号1、2、3の部分木が有する“レタス、キャベツ、白菜”を2段上位の“農作物”に変換したものであるため、そのスコア4 fは $12 \times 0.7 \times 0.7 = 5.88$ となる。

【0081】次に、部分木判定部1 eは、各部分木をスコアでソートし、スコアの高い部分木から要約を出力すべきかどうかの判定を行う。ここで、最もスコアの高い部分木番号1の部分木(レタス-の-価格-が-高い)はリンク先がないので(ステップS1)、部分木番号1の部分木の判定4 gに1を設定する(ステップS2)。

【0082】次にスコアの高い部分木番号4の部分木(野菜-の-価格-が-高い)はリンク先を有し(ステップS1)、リンク先である部分木番号1、2、3の部分木の補正頻度4 dが、部分木番号4の部分木の補正頻度4 d以上ではないが(ステップS3)、リンク先である部分木番号1の部分木の判定4 gが1であるため(ステップS4)、部分木番号4の部分木を要約として出力しないこととし、その判定4 gを-1とする(ステップS5)。

【0083】次にスコアの高い部分木番号5の部分木(農作物-の-価格-が-高い)はリンク先を有し(ステップS1)、リンク先である部分木番号4の部分木(野菜-の-価格-が-高い)の補正頻度4 d(12)が部分木番号5の部分木の補正頻度4 d(12)以上なので(ステップS3)、部分木番号5の部分木からは要約を生成しないこととし、判定4 gに-1を設定する(ステップS5)。

【0084】次にスコアの高い部分木番号2、3の部分木はリンク先を有していないため(ステップS1)、それらの判定4 gに1を設定する(ステップS2)。以上により、図17に示すような判定情報が生成される。部分木判定部1 eは、このように生成された判定情報を句合成部1 fに送る。

【0085】判定情報を受け取った句合成部1 fは、図5のフローチャートに従って要約の合成を行う。ここで、部分木番号1の部分木は判定4 gが1なので(ステ

ップS21)、要約「レタスの価格が高い」を合成し、補正頻度[10]とともに表示する(ステップS22)。また、部分木番号1の部分木はリンク先を有していないため(ステップS24)、次の部分木の判定に移る。

【0086】部分木番号4、5の部分木は判定-1なので(ステップS21)、スキップする(ステップS23)。部分木番号2、3の部分木の判定4 gは1なので(ステップS21)、要約「キャベツの価格が高い、白菜の価格が高い」を合成し、頻度[1]とともに表示する(ステップS22)。

【0087】以上より、図18に示す要約が生成される。このように本形態では、文解析部1 aによって、入力文書の解析木を生成し、部分木作成部1 bによって、解析木から部分木を抽出し、部分木変換リンク部1 cで部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木判定部1 eは、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否かを判定し、句合成部1 fによって、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示することとしたため、頻度が高く、抽象化された「野菜の価格が高い」のスコアを上回った「レタスの価格が高い」を要約として選択することができる。

【0088】なお、上記の処理機能は、コンピュータによって実現することができる。その場合、文書要約装置が有すべき機能の処理内容は、コンピュータで読み取り可能な記録媒体に記録されたプログラムに記述しておく。そして、このプログラムをコンピュータで実行することにより、上記処理がコンピュータで実現される。コンピュータで読み取り可能な記録媒体としては、磁気記録装置や半導体メモリ等がある。市場に流通させる場合には、CD-ROM(Compact Disk Read Only Memory)やフロッピーディスク等の可搬型記録媒体にプログラムを格納して流通させたり、ネットワークを介して接続されたコンピュータの記憶装置に格納しておき、ネットワークを通じて他のコンピュータに転送することもできる。コンピュータで実行する際には、コンピュータ内のハードディスク装置等にプログラムを格納しておき、メインメモリにロードして実行する。

【0089】

【発明の効果】以上説明したように本発明の文書要約装置では、文解析手段によって、入力文書の解析木を生成し、部分木作成手段によって、解析木から部分木を抽出し、部分木変換リンク手段で部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木判定手段によって、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否かを判定し、句合成部手段によって、出力すると判定した部分木を、出力すると判定され

た部分木の変換元の部分木に対応付けて表示することとした為、抽象化が進んだ要約であっても、同時に対応付けて表示される変換元の要約を参照できる為、要約の内容把握が容易になる。

【0090】また、本発明の文書要約装置では、判定を行う部分木の補正頻度が、判定を行う部分木の変換元に該当する部分木の補正頻度よりも低い場合、部分木の出力を行わない判定を行うこととした為、不適切な要約を排除することが可能となり、要約の内容把握が容易になる。

【0091】さらに、本発明の文書要約装置では、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示することとした為、関連する要約をまとめて表示することが可能となる。

【0092】また、本発明の文書要約方法では、入力文書の解析木を生成し、解析木から部分木を抽出し、部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否かを判定し、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示することとした為、抽象化が進んだ要約であっても、同時に対応付けて表示される変換元の要約を参照できる為、要約の内容把握が容易になる。

【0093】さらに、本発明の記録媒体では、入力文書の解析木を生成し、解析木から部分木を抽出し、部分木を変換して新たな部分木を生成し、生成した部分木及びその変換元との関係を示すポイントを付与し、部分木にスコアを付与し、スコアが高い順番に部分木を要約として出力するか否かを判定し、出力すると判定した部分木を、出力すると判定された部分木の変換元の部分木に対応付けて表示する機能をコンピュータに行わせるプログラムを記録することとした為、当該プログラムをコンピュータ上で起動させることにより、抽象化が進んだ要約であっても、同時に対応付けて表示される変換元の要約を参照できる為、要約の内容把握が容易になる。

【図面の簡単な説明】

【図1】 文書要約装置の構成図である。

【図2】 シソーラスに記録されているシソーラス辞書

の内容を示した図である。

【図3】 部分木判定部が行う判定の判定手順を例示したフローチャートである。

【図4】 部分木判定部が行う判定の判定手順を例示したフローチャートである。

【図5】 要約の合成手順を例示したフローチャートである。

【図6】 複数文書を入力文書とした場合における文書要約装置の要約生成過程を示した図である。

【図7】 部分木変換リンク部によって生成された結果を示した図である。

【図8】 部分木判定部が生成した判定情報を示した図である。

【図9】 生成される要約を示した図である。

【図10】 複数文書を入力文書とした場合における文書要約装置の要約生成過程を示した図である。

【図11】 部分木変換リンク部によって生成された結果を示した図である。

【図12】 部分木判定部が生成した判定情報を示した図である。

【図13】 生成される要約を示した図である。

【図14】 複数文書を入力文書とした場合における文書要約装置の要約生成過程を示した図である。

【図15】 部分木変換リンク部によって生成された結果を示した図である。

【図16】 部分木判定部が付与したスコアを示す図である。

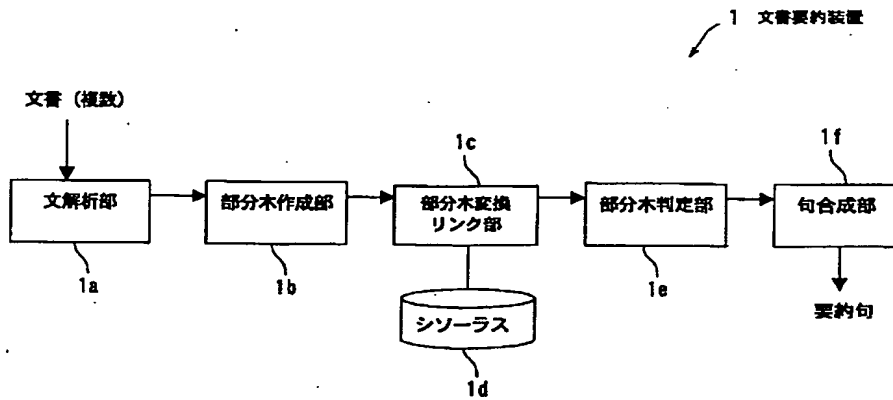
【図17】 部分木判定部が生成した判定情報を示した図である。

【図18】 生成される要約を示した図である。

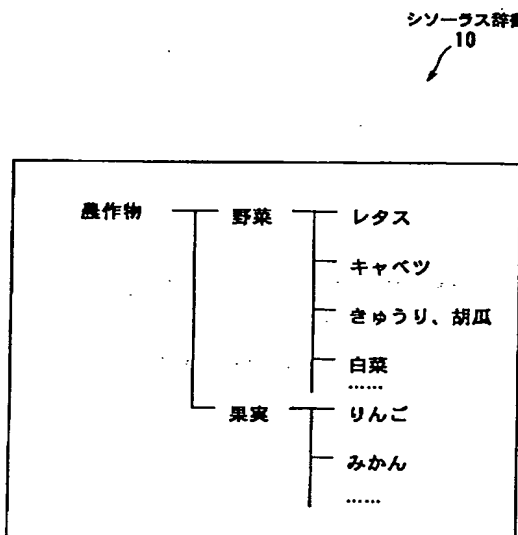
【符号の説明】

- 1 文書要約装置
- 1 a 文解析部
- 1 b 部分木作成部
- 1 c 部分木変換リンク部
- 1 d シソーラス
- 1 e 部分木判定部
- 1 f 句合成部
- 1 0 シソーラス辞書

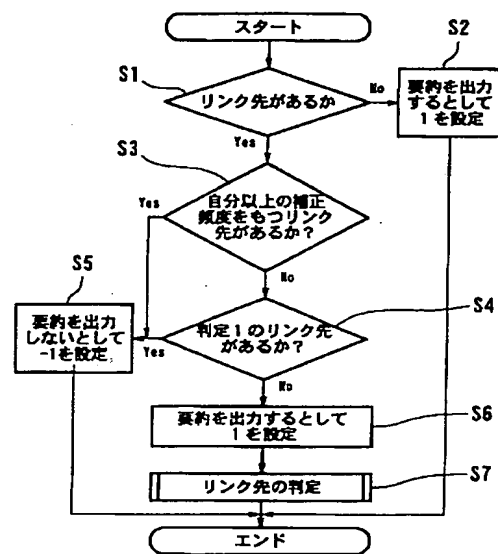
【図1】



【図2】



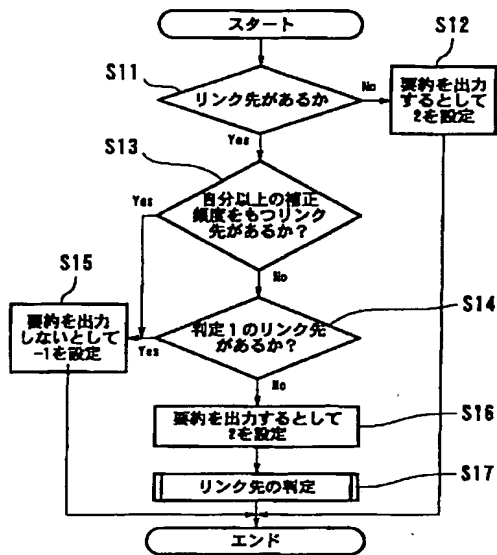
【図3】



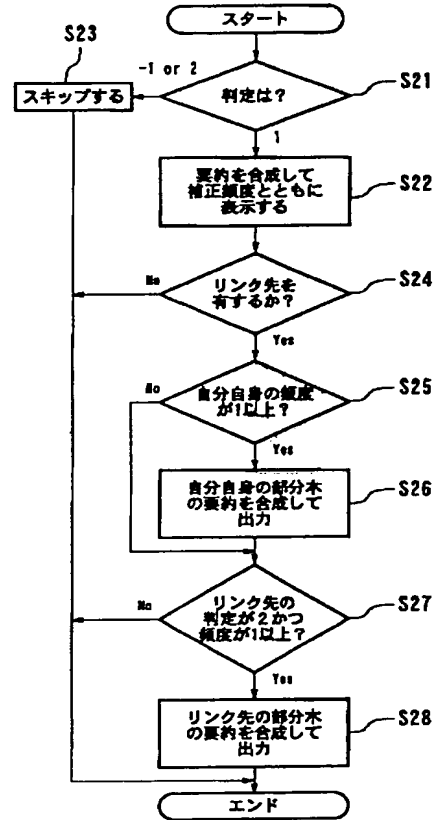
【図11】

部分木番号	変換部分木情報	頻度	補正頻度	リンク
1	野菜 - の - 価格 - が - 高い	2	2	-1
2	野菜 - の - 価格 - の - 全国調査	1	1	-1
3	農作物 - の - 価格 - を - 調査	1	2	2
4	農作物 - の - 価格 - が - 高い	0	2	1
5	農作物 - の - 価格 - の - 全国調査	0	1	2
6	野菜 - の - 価格 - の - 調査	0	1	2

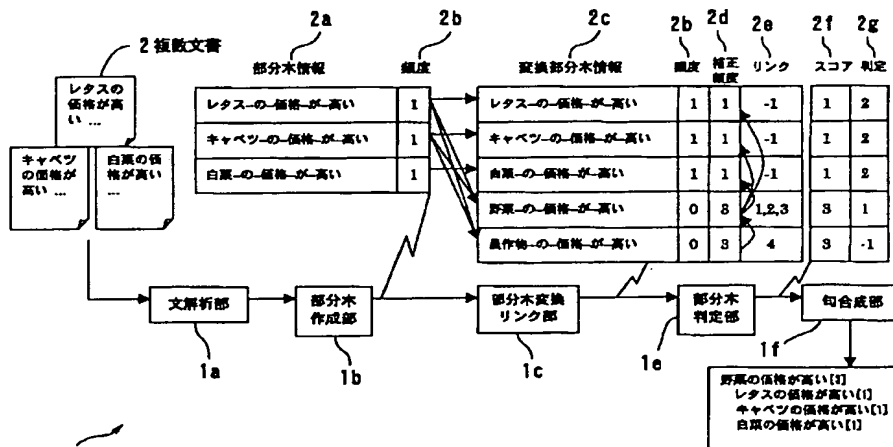
【図4】



【図5】



【図6】



【図7】

部分木番号	交換部分木情報	頻度	補正頻度	リンク
1	レタス-の-価格-が-高い	1	1	-1
2	キャベツ-の-価格-が-高い	1	1	-1
3	白菜-の-価格-が-高い	1	1	-1
4	野菜-の-価格-が-高い	0	3	1, 2, 3
5	農作物-の-価格-が-高い	0	3	4

【図8】

部分木番号	交換部分木情報	頻度	補正頻度	リンク	スコア	判定
4	野菜-の-価格-が-高い	0	3	1, 2, 3	3	1
5	農作物-の-価格-が-高い	0	3	4	3	-1
1	レタス-の-価格-が-高い	1	1	-1	1	2
2	キャベツ-の-価格-が-高い	1	1	-1	1	2
3	白菜-の-価格-が-高い	1	1	-1	1	2

【図9】

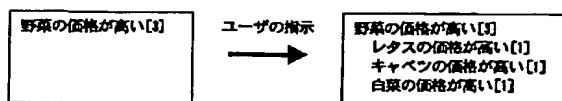
(a)

野菜の価格が高い[3]
レタスの価格が高い[1]
キャベツの価格が高い[1]
白菜の価格が高い[1]

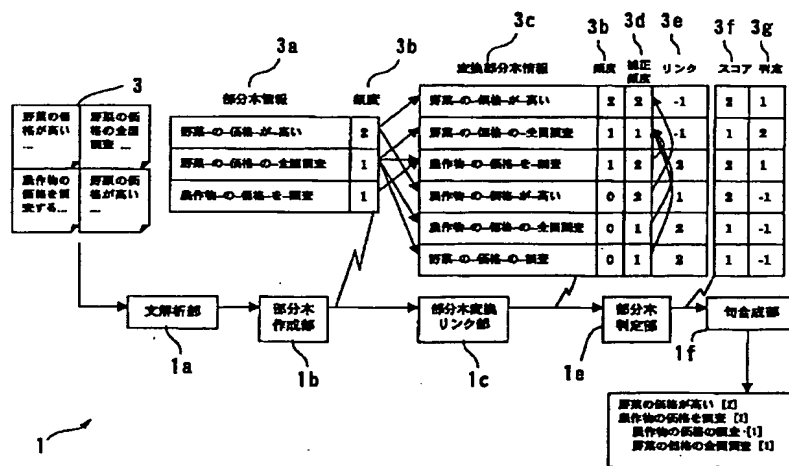
【図18】

レタスの価格が高い[10]
キャベツの価格が高い[1]
白菜の価格が高い[1]

(b)



【図10】



【図12】

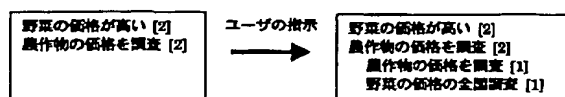
部分木番号	交換部分木情報	頻度	補正頻度	リンク	スコア	判定
1	野菜-の-価格-が-高い	2	2	-1	2	1
3	農作物-の-価格-を-調査	1	2	2	2	1
4	農作物-の-価格-が-高い	0	2	1	2	-1
2	野菜-の-価格-の-全国調査	1	1	-1	1	2
5	農作物-の-価格-の-全国調査	0	1	2	1	-1
6	野菜-の-価格-の-調査	0	1	2	1	-1

【図13】

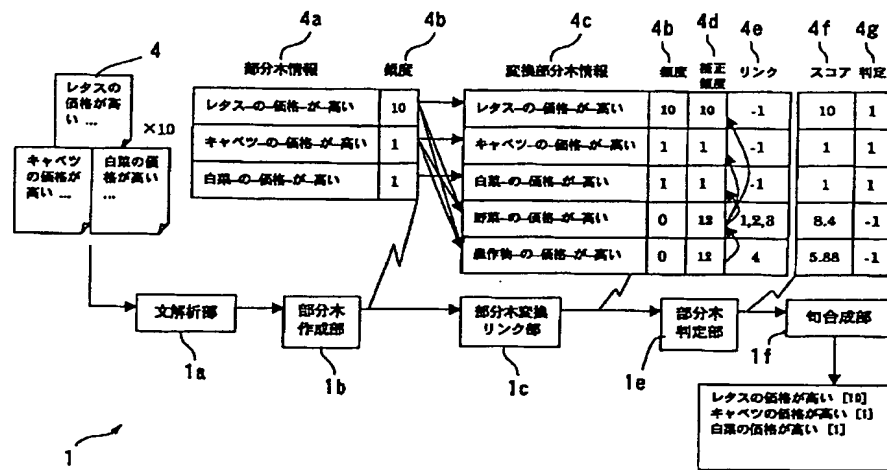
(a)

野菜の価格が高い [2]
農作物の価格を調査 [2]
農作物の価格を調査 [1]
野菜の価格の全国調査 [1]

(b)



【図14】



【図15】

部分木番号	変換部分木情報	頻度	補正頻度	リンク
1	レタス - の - 価格 - が - 高い	10	10	-1
2	キャベツ - の - 価格 - が - 高い	1	1	-1
3	白菜 - の - 価格 - が - 高い	1	1	-1
4	野菜 - の - 価格 - が - 高い	0	12	1, 2, 3
5	農作物 - の - 価格 - が - 高い	0	12	4

【図16】

部分木番号	変換部分木情報	頻度	補正頻度	リンク	スコア
1	レタス - の - 価格 - が - 高い	10	10	-1	10
2	キャベツ - の - 価格 - が - 高い	1	1	-1	1
3	白菜 - の - 価格 - が - 高い	1	1	-1	1
4	野菜 - の - 価格 - が - 高い	0	12	1, 2, 3	8.4
5	農作物 - の - 価格 - が - 高い	0	12	4	5.88

【 図 17 】

部分木 番号		4c 交換部分木情報	4b 頻度	4d 補 正 頻度	4e リンク	4f スコア	4g 判定
1		レタス - の - 価 格 - が - 高 い	10	10	-1	10	1
4		野菜 - の - 価 格 - が - 高 い	0	12	1, 2, 3	8.4	-1
5		農作物 - の - 価 格 - が - 高 い	0	12	4	5.88	-1
2		キャベツ - の - 価 格 - が - 高 い	1	1	-1	1	1
3		白菜 - の - 価 格 - が - 高 い	1	1	-1	1	1

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.